Electronic Journal of Statistics
Vol. 13 (2019) 4015–4034
ISSN: 1935-7524
https://doi.org/10.1214/19-EJS1622

Why scoring functions cannot assess tail properties

Jonas R. Brehmer

Institute of Mathematics
University of Mannheim, Germany
e-mail: jbrehmer@mail.uni-mannheim.de

and

Kirstin Strokorb

School of Mathematics
Cardiff University, United Kingdom
e-mail: strokorbk@cardiff.ac.uk

Abstract: Motivated by the growing interest in sound forecast evaluation techniques with an emphasis on distribution tails rather than average behaviour, we investigate a fundamental question arising in this context: Can statistical features of distribution tails be elicitable, i.e. be the unique minimizer of an expected score? We demonstrate that expected scores are not suitable to distinguish genuine tail properties in a very strong sense. Specifically, we introduce the class of max-functionals, which contains key characteristics from extreme value theory, for instance the extreme value index. We show that its members fail to be elicitable and that their elicitation complexity is in fact infinite under mild regularity assumptions. Further we prove that, even if the information of a max-functional is reported via the entire distribution function, a proper scoring rule cannot separate max-functional values. These findings highlight the caution needed in forecast evaluation and statistical inference if relevant information is encoded by such functionals.

MSC 2010 subject classifications: Primary 62C05, 62G32; secondary 91B06.
Keywords and phrases: Elicitability, elicitation complexity, extreme value index, max-functional, proper scoring rule, scoring functions, consistency, tail equivalence.

Received May 2019.

1. Introduction

Many of our day-to-day decisions rely on our ability to produce reasonable forecasts for quantities of interest. For example, production planning involves forecasts on consumer demand, decisions in farming depend on information about the likely weather conditions and financial risk management uses statistical features of portfolio losses. Usually, such quantities are modelled via a random variable Y having an unknown probability distribution and the reasonable actions

of a decision maker depend on the properties of this distribution. Forecasts can encode such properties via real numbers, e.g. means or quantiles of the distribution, via sets, e.g. a confidence interval, or by a report of the whole distribution function.

When several competing forecasts are available, a crucial problem is to determine which one is most valuable. A principled approach to this task is to compare the forecasts to a set of realizations of Y via a scoring rule or a scoring function, see e.g. Gneiting and Raftery (2007) and Gneiting (2011). A *scoring function* assigns a real-valued score based on a forecast and a realizing observation. If a functional, i.e. a statistical property, of a distribution is the unique minimizer of the expected score with respect to this distribution, it is called *elicitable*. Elicitability is a desirable property for comparative forecast evaluation, where it can be used to incentivize risk-neutral forecasters to report their beliefs (Gneiting, 2011). Moreover, elicitable functionals enable regression and M-estimation (Fissler and Ziegel, 2016; Gneiting, 2011) and are central to various machine learning algorithms (Steinwart et al., 2014; Frongillo and Kash, 2018). Recent theoretical advances on scoring functions and elicitability in the real-valued case can be found in Lambert, Pennock and Shoham (2008), Gneiting (2011) and Steinwart et al. (2014). More general vector-valued functionals are treated in Frongillo and Kash (2015, 2018) and Fissler and Ziegel (2016, 2019).

Many statistical functionals such as expectations, quantiles, and expectiles are elicitable and there exist convenient characterizations of the corresponding classes of consistent scoring functions, cf. Gneiting (2011) and the references therein. On the other hand, several widely considered functionals fail to be elicitable, for instance the variance, the mode (Heinrich, 2014) and the prominent financial risk measure Expected Shortfall (ES) (Weber, 2006; Gneiting, 2011). The non-elicitability of the latter functional can be addressed via more general notions of elicitability: Fissler and Ziegel (2016) show that ES is *jointly elicitable* with the risk measure Value at Risk (VaR), where the latter is simply an extreme quantile. In other words, ES has *elicitation complexity* equal to two in the sense of Frongillo and Kash (2018). In this particular instance the elicitability problems associated with ES can be resolved, at the cost of considering a higher dimensional problem.

More generally, there is a recent growing interest in sound forecast evaluation techniques with an emphasis on distribution tails rather than average behaviour. For instance, Friederichs and Thorarinsdottir (2012) investigate the use of scoring rules for distribution classes central to extreme value theory, and Diks, Panchenko and van Dijk (2011), Lerch et al. (2017) as well as Holzmann and Klar (2017) consider weighted scoring rules for forecasts of distribution tails. An event-based approach to evaluate whether exceedances of high thresholds are predicted correctly is pursued by Stephenson et al. (2008) and Ferro and Stephenson (2011). Closely connected is the verification tool of Taillardat et al. (2019) which is based on the asymptotic behavior of the continuously ranked probability score (CRPS), conditional on high realizations. A fundamental question arising in this context is to what extent, and in which sense,

statistical features of distribution tails are elicitable. The latter problem is the central theme of this manuscript.

In our approach to this question we introduce the concept of max-functionals which naturally arises from a key feature shared by the statistical functionals that are typically considered in extreme value theory. We demonstrate that max-functionals fail to be elicitable in a very strong sense. Consequently, it is natural to ask whether part of the problem can be mitigated by abandoning point forecasts in favor of reports of the entire distribution function. In this regard we generalize a result by Taillardat et al. (2019) and show that it is an inherent property of *all* proper scoring rules that they cannot perfectly distinguish among different max-functional values.

The manuscript is organized as follows. In Section 2 we review the three notions of elicitability that are used in the recent literature. Section 3 introduces the class of max-functionals and shows that they cannot be elicitable and that their elicitation complexity is infinite under mild assumptions. Section 4 provides examples of widely used max-functionals. In Section 5 we turn to reports of entire distributions. We show that arbitrary large differences in tail behaviour, either quantified by tail equivalence or max-functionals, can remain undetected by proper scoring rules. Section 6 concludes with a discussion of the results.

2. Prerequisites: elicitability and elicitation complexity

For the reader's convenience this section recalls the central definitions of elicitability and reviews basic findings. A more detailed overview of the existing literature is given in Fissler and Ziegel (2016) and Gneiting (2011), whose notation we follow here. Let $\mathsf{O} \subseteq \mathbb{R}^d$ be a fixed set, called *observation domain*, equipped with Borel σ-algebra $\mathcal{O}$. We use $\mathcal{F}$ to denote a collection of probability distributions on $(\mathsf{O}, \mathcal{O})$, whilst also identifying probability distributions with their cumulative distribution functions. A *functional* will be a mapping $T : \mathcal{F} \to \mathsf{A}$ where $\mathsf{A} \subseteq \mathbb{R}^n$ is called *action domain*. A measurable function $g : \mathsf{O} \to \mathbb{R}$ is called *$\mathcal{F}$-integrable* if it is integrable with respect to all $F \in \mathcal{F}$. Analogously, a function $g : \mathsf{A} \times \mathsf{O} \to \mathbb{R}$ is called $\mathcal{F}$-integrable if for all $x \in \mathsf{A}$ the function $y \mapsto g(x, y)$ is integrable with respect to all $F \in \mathcal{F}$. We use the short notation

$$\bar{h}(F) := \int_{\mathsf{O}} h(y)\,\mathrm{d}F(y) \quad \text{and} \quad \bar{g}(x, F) := \int_{\mathsf{O}} g(x, y)\,\mathrm{d}F(y)$$

for $\mathcal{F}$-integrable functions h, g and $x \in \mathsf{A}$, $F \in \mathcal{F}$.

Scoring functions and elicitability In the following, $S : \mathsf{A} \times \mathsf{O} \to \mathbb{R}$ denotes a *scoring function*, i.e. an $\mathcal{F}$-integrable function. The central concepts connecting scoring functions and statistical functionals are consistency and elicitability.

Definition 2.1 (Consistency)**.** A scoring function $S : \mathsf{A} \times \mathsf{O} \to \mathbb{R}$ is *$\mathcal{F}$-consistent* for a functional $T : \mathcal{F} \to \mathsf{A}$ if for all $x \in \mathsf{A}$ and $F \in \mathcal{F}$ we have $\bar{S}(x, F) \geq \bar{S}(T(F), F)$. It is called *strictly $\mathcal{F}$-consistent* for T if it is $\mathcal{F}$-consistent for T and for all $x \in \mathsf{A}$ and $F \in \mathcal{F}$ the equality $\bar{S}(x, F) = \bar{S}(T(F), F)$ implies $x = T(F)$.

Definition 2.2 ((Joint) elicitability)**.** A functional $T : \mathcal{F} \to \mathsf{A} \subseteq \mathbb{R}^n$ is called *elicitable* if there exists a strictly $\mathcal{F}$-consistent scoring function for T. It is called *jointly elicitable with the functional* $T' : \mathcal{F} \to \mathsf{A}' \subseteq \mathbb{R}^k$ if (T, T') is an elicitable functional.

An important necessary condition that a statistical functional needs to satisfy in order to be elicitable is *convexity of level sets*, which goes back to Osband (1985), cf. for instance Gneiting (2011, Theorem 6) and Lambert, Pennock and Shoham (2008, Lemma 1) for a proof.

Theorem 2.3 (Convexity of level sets)**.** *Let* $T : \mathcal{F} \to \mathsf{A}$ *be an elicitable functional. If* $F_0, F_1 \in \mathcal{F}$ *and* $\lambda \in (0,1)$ *are such that* $F_\lambda = \lambda F_1 + (1-\lambda)F_0 \in \mathcal{F}$*, then* $T(F_0) = T(F_1) = t$ *implies* $T(F_\lambda) = t$*.*

Example 2.4. The simplest example of an elicitable functional is the mean of a distribution. More precisely, let $g : \mathsf{O} \to \mathbb{R}$ be such that g and g^2 are $\mathcal{F}$-integrable and define $T : \mathcal{F} \to \mathbb{R}$ via $T(F) = \bar{g}(F)$. Then T is elicitable with a strictly $\mathcal{F}$-consistent scoring function given by $S(x, y) = (x - g(y))^2$, the ubiquitous squared error loss. Likewise, the moment functionals defined via $T_k(F) := \int y^k \, \mathrm{d}F(y)$ for $k \in \mathbb{N}$ are elicitable.

A simple example of a non-elicitable functional is the variance functional $T_{\mathrm{var}}(F) := T_2(F) - T_1(F)^2$, whose non-elicitability follows directly from Theorem 2.3. Nevertheless, T_{var} is jointly elicitable since the vector (T_1, T_{var}) can be obtained from the elicitable vector (T_1, T_2) via a bijection and hence it is elicitable, see e.g. Gneiting (2011, Theorem 4). Another notable property is that on every subset of $\mathcal{F}$ where T_1 is constant, T_{var} reduces to a shifted version of the second moment T_2 and is thus elicitable on this subset. That is, T_{var} is conditionally elicitable given T_1 in the following sense.

Definition 2.5 (Conditional elicitability)**.** Let $T : \mathcal{F} \to \mathsf{A} \subseteq \mathbb{R}^n$ and $T' : \mathcal{F} \to \mathsf{A}' \subseteq \mathbb{R}^k$ be functionals and let T' be elicitable. For any $x \in \mathsf{A}'$ define the set

$$\mathcal{F}_x := \{F \in \mathcal{F} \mid T'(F) = x\}.$$

Then the functional T is called *conditionally elicitable given* T' if for any $x \in \mathsf{A}'$ its restriction to the class $\mathcal{F}_x$ is elicitable.

The concept of conditional elicitability was first introduced by Emmer, Kratz and Tasche (2015) and motivated by a conditional backtesting approach for Expected Shortfall (ES) forecasts. A slight generalization was given by Fissler and Ziegel (2016). Our definition coincides with the one from Fissler and Ziegel (2016) except that we drop the condition that T' has elicitable components and only require it to be elicitable. This allows for a more convenient presentation of our results below.

Neither joint elicitability nor conditional elicitability imply elicitability, which follows from Example 2.4 with the variance functional serving as a counterexample. If a functional T is jointly elicitable with the functional T', and T' is elicitable, then it is conditionally elicitable given T'. Conversely, as discussed in Fissler and Ziegel (2016), it is unclear under which conditions a conditionally elicitable functional is jointly elicitable.

Elicitation complexity The definitions of joint elicitability and conditional elicitability both require a second elicitable functional T' accompanying the functional of interest. The distinction between both functionals is made more explicit in the concept of *elicitation complexity*. To illustrate this, recall Example 2.4 and note that the variance functional satisfies $T_{\text{var}} = f(T_1, T_2)$, where $f(x_1, x_2) = x_2 - x_1^2$. Since T_1 and T_2 are elicitable, we say that the variance functional has complexity 2. In general, T has elicitation complexity at most k if there is an elicitable functional $T' : \mathcal{F} \to \mathsf{A}' \subseteq \mathbb{R}^k$ such that $T = f(T')$ holds. Any f and T' satisfying this condition are then called *link function* and *intermediate functional*, respectively. The smallest dimension k for which such a representation is feasible is the elicitation complexity.

Definition 2.6 (Elicitation complexity)**.** For any set of distribution functions $\mathcal{F}$ the set of $\mathbb{R}^k$-valued elicitable functionals defined on $\mathcal{F}$ is denoted via $\mathcal{E}_k(\mathcal{F})$. For a functional $T : \mathcal{F} \to \mathsf{A} \subseteq \mathbb{R}$ and sets $\mathcal{C}_k \subseteq \mathcal{E}_k(\mathcal{F})$ the *elicitation complexity of* T *with respect to* $(\mathcal{C}_k)_{k\in\mathbb{N}}$ is defined via

$$\mathsf{elic}(T) := \min\{k \in \mathbb{N} \mid \exists\, T' \in \mathcal{C}_k : T = f \circ T' \text{ for some } f : T'(\mathcal{F}) \to \mathsf{A}\}.$$

If the minimum is not attained for any $k \in \mathbb{N}$, the elicitation complexity of T with respect to $(\mathcal{C}_k)_{k\in\mathbb{N}}$ is infinite and we write $\mathsf{elic}(T) = \infty$.

Elicitation complexity was introduced by Lambert, Pennock and Shoham (2008) and further analyzed in Frongillo and Kash (2018), the latter motivated by its role in empirical risk minimization (ERM) algorithms in machine learning. Intuitively speaking, it replaces the question *whether* a functional is elicitable by the question *how complex* it is to elicit the functional.

If no regularity conditions are imposed on f or T', this can lead to small complexities without clear benefits in applications. More precisely, if f is arbitrary and $\mathcal{C}_k = \mathcal{E}_k(\mathcal{F})$ is chosen, pathological choices of f, like bijections from $\mathbb{R}^k$ to $\mathbb{R}$, cause all functionals to have complexity 1, as demonstrated by Frongillo and Kash (2018, Remark 2). To avoid such problems, it is standard to choose suitable subclasses $\mathcal{C}_k$ of intermediate functionals. One possible choice, which is used by Frongillo and Kash (2018) as well as Dearborn and Frongillo (2019), is $\mathcal{C}_k := \mathcal{I}_k(\mathcal{F}) \cap \mathcal{E}_k(\mathcal{F})$, where $\mathcal{I}_k(\mathcal{F})$ is the set of $\mathbb{R}^k$-valued identifiable functionals on $\mathcal{F}$. Another possibility, implicitly used by Lambert, Pennock and Shoham (2008) is to define $\mathcal{C}_k$ to be a subclass of all functionals which have elicitable components.

Lastly, it is also possible to impose regularity on the link function f, e.g. by requiring differentiability or continuity. Notably, joint elicitability can be understood as a version of elicitation complexity where the link function is the projection on the last component (Frongillo and Kash, 2018).

We need to be cautious when interpreting elicitation complexity, since imposing different regularity conditions via $(\mathcal{C}_k)_{k\in\mathbb{N}}$ can lead to different elicitation complexities for the same functional, see Frongillo and Kash (2018, Subsection 2.2) for an example. In particular, some $\mathbb{R}^k$-valued functional might be elicitable and simultaneously have elicitation complexity strictly greater than

k. Conversely, a functional can have elicitation complexity 1, although it is not itself elicitable, as illustrated in Frongillo and Kash (2018, Remark 1).

We conclude this section with a lemma which considers the properties of a functional T if it is restricted to some subclass $\mathcal{F}_2 \subseteq \mathcal{F}$. The first statement corresponds to the first part of Lemma 2.11 of Fissler and Ziegel (2015), the second and third statement are simple extensions. Their proofs are straightforward and therefore omitted.

Lemma 2.7. *Let $T : \mathcal{F} \to \mathsf{A}$ be a functional and let $\mathcal{F}_2 \subseteq \mathcal{F}$ be non-empty.*

(a) If T is elicitable, then the restricted functional $T_{|\mathcal{F}_2}$ is elicitable.
(b) If $\mathsf{elic}(T) = k$ with respect to $(\mathcal{C}_k)_{k\in\mathbb{N}}$ and we define $\mathcal{C}_k^2 := \{T'_{|\mathcal{F}_2} \mid T' \in \mathcal{C}_k\}$, then $\mathsf{elic}(T_{|\mathcal{F}_2}) \leq k$ with respect to $(\mathcal{C}_k^2)_{k\in\mathbb{N}}$.
(c) If $\mathsf{elic}(T) = k$ with respect to $(\mathcal{C}_k)_{k\in\mathbb{N}}$ and sets $(\mathcal{C}'_k)_{k\in\mathbb{N}}$ satisfy $\mathcal{C}_k \subseteq \mathcal{C}'_k$ for all $k \in \mathbb{N}$, then $\mathsf{elic}(T) \leq k$ with respect to $(\mathcal{C}'_k)_{k\in\mathbb{N}}$.

3. The elicitation complexity of max-functionals

This section introduces max-functionals, the central objects of our study, and investigates their elicitability as well as their elicitation complexity. Henceforth, let $\mathcal{F}$ always denote a *convex* class of distributions.

Definition 3.1. A functional $T : \mathcal{F} \to \mathbb{R}$ is called *max-functional* if

$$T(\lambda F_1 + (1-\lambda)F_0) = \max(T(F_0), T(F_1))$$

holds for all $F_0, F_1 \in \mathcal{F}$ and $\lambda \in (0,1)$.

The essential feature of a max-functional is that its value on convex combinations of distributions is determined by the values attained on the extreme points. Equivalently, we can also define min-functionals and all results carry over with minor modifications. The constant functional is the simplest max-functional, but we will usually not be interested in this trivial case. Instead, Section 4 collects some non-trivial examples of max-functionals that are routinely considered in extreme value theory. Also note that, by definition, restrictions of max-functionals to a certain set of values are again max-functionals.

Lemma 3.2. *Let $T : \mathcal{F} \to \mathbb{R}$ be a max-functional and $A \subset \mathbb{R}$ a set. Set $\mathcal{F}_A := \{F \in \mathcal{F} \mid T(F) \in A\}$, then $\mathcal{F}_A$ is convex and the restricted functional $T : \mathcal{F}_A \to A \subset \mathbb{R}$ is also a max-functional.*

Non-elicitability of max-functionals We start by proving that max-functionals cannot be elicitable. As remarked in Section 2 the usual way to show that a functional is not elicitable consists of applying Theorem 2.3, i.e. showing that it fails to have convex level sets. However, any max-functional has convex level sets by definition. So this approach is not feasible, as in the case of the mode functional (Heinrich, 2014). Instead, we employ the following new criterion.

Theorem 3.3. *Let $T : \mathcal{F} \to \mathsf{A}$ be a functional. If there are $F_0, F_1 \in \mathcal{F}$ such that $T(F_0) \neq T(F_1)$ and*

$$T(\lambda F_1 + (1-\lambda)F_0) \in \{T(F_0), T(F_1)\} \quad \textit{for all } \lambda \in (0,1),$$

then T is not elicitable.

Proof. Set $x_0 := T(F_0)$, $x_1 := T(F_1)$ and $F_\lambda := \lambda F_1 + (1-\lambda)F_0$ and let $x_0 \neq x_1$. Assume that S is a strictly consistent scoring function for T. Then we have

$$\begin{aligned}\bar{S}(x_0, F_\lambda) - \bar{S}(x_1, F_\lambda) = \lambda(\bar{S}(x_0, F_1) - \bar{S}(x_1, F_1)) \\ + (1-\lambda)(\bar{S}(x_0, F_0) - \bar{S}(x_1, F_0))\end{aligned}$$

and the first difference $\bar{S}(x_0, F_1) - \bar{S}(x_1, F_1)$ is positive, while the second difference $\bar{S}(x_0, F_0) - \bar{S}(x_1, F_0)$ is negative. Consequently, $\bar{S}(x_0, F_\lambda) = \bar{S}(x_1, F_\lambda)$ for some $\lambda \in (0,1)$. Since either $T(F_\lambda) = x_0$ or $T(F_\lambda) = x_1$ holds by assumption, we arrive at a contradiction. □

Corollary 3.4. *If $T : \mathcal{F} \to \mathbb{R}$ is a non-constant max-functional, then it is not elicitable.*

Loosely speaking, Theorem 3.3 states that elicitable functionals cannot be piecewise constant on convex combinations of distributions. It is closely connected to Theorem 2.3, but of independent interest beyond its use to establish non-elicitability for max-functionals. Fissler, Hlavinová and Rudloff (2019) use similar arguments as in the proof of Theorem 3.3 to study necessary conditions for the level sets of T in the context of set-valued functionals $T : \mathcal{F} \to 2^{\mathsf{A}}$, where 2^{A} denotes the power set of A. Apart from that Frongillo and Kash (2018) state that '*no nonconstant finite-valued property is identifiable*'. Theorem 3.3 implies the following analogon.

Corollary 3.5. *If $T : \mathcal{F} \to \mathbb{R}$ is a non-constant finite-valued functional, then it is not elicitable.*

Elicitation complexity of max-functionals Turning from the elicitability question to the elicitation complexity of max-functionals, the question of elicitation complexity is only meaningful in relation to a family of sets $(\mathcal{C}_k)_{k\in\mathbb{N}}$, where each set $\mathcal{C}_k \subset \mathcal{E}_k(\mathcal{F})$ is a collection of reasonably regular $\mathbb{R}^k$-valued elicitable functionals, cf. Section 2. Our major regularity requirement is *mixture-continuity* as in Bellini and Bignozzi (2015) and Fissler and Ziegel (2019).

Definition 3.6. A functional $T : \mathcal{F} \to \mathsf{A}$ is called *mixture-continuous* if for all $F_0, F_1 \in \mathcal{F}$ such that $\lambda F_1 + (1-\lambda)F_0 \in \mathcal{F}$ for all $\lambda \in [0,1]$, the mapping

$$[0,1] \to \mathsf{A}, \quad \lambda \mapsto T(\lambda F_1 + (1-\lambda)F_0)$$

is a continuous function.

Many statistical properties are mixture-continuous, e.g. ratios of expectations, quantiles and expectiles, see Fissler and Ziegel (2019) for details. Lambert, Pennock and Shoham (2008) consider only continuous functionals and

Fissler and Ziegel (2019) and Bellini and Bignozzi (2015) show that under weak assumptions, an elicitable functional T' is mixture-continuous if its expected score function $x \mapsto \bar{S}(x, F)$ is continuous for all $F \in \mathcal{F}$. Therefore, a functional which is not mixture-continuous can have discontinuous expected scores, leading to difficulties in forecast evaluation, estimation and regression.

To avoid further degenerate behaviour, we impose a richness assumption on potential intermediate functionals T' in the sense that we require the image $T'(\mathcal{F}) \subseteq \mathbb{R}^k$ to have at least non-empty interior. This assumption is natural for large enough classes $\mathcal{F}$ and was, for instance, used by Fissler and Ziegel (2016, 2019) when establishing results on consistent scoring functions for T'.

In addition to mixture continuity, we follow Lambert, Pennock and Shoham (2008) and consider only functionals with elicitable components. Summarising, the first family of functionals which we consider in our complexity result is

$$\mathcal{U}_k := \left\{ T' \in \mathcal{E}_k(\mathcal{F}) \;\middle|\; \begin{array}{l} T' \text{ mixture-continuous with elicitable} \\ \text{components, } \operatorname{int}(T'(\mathcal{F})) \neq \emptyset \end{array} \right\},$$

where $\operatorname{int}(B)$ denotes the interior of a set $B \subseteq \mathbb{R}^k$. Alternatively, we require that the image $T'(\mathcal{F})$ of a potential intermediate functional T' has not only non-empty interior, but is itself an open set, i.e. we consider the family

$$\mathcal{V}_k := \left\{ T' \in \mathcal{E}_k(\mathcal{F}) \;\middle|\; \begin{array}{l} T' \text{ mixture-continuous with elicitable} \\ \text{components, } T'(\mathcal{F}) \text{ open} \end{array} \right\}.$$

We are now in position to consider the elicitation complexity of max-functionals with respect to these families.

Theorem 3.7. *Let $T : \mathcal{F} \to \mathbb{R}$ be a max-functional. Then the following hold true.*

(a) T has elicitation complexity ∞ with respect to $(\mathcal{U}_k)_{k\in\mathbb{N}}$ unless $T(\mathcal{F})$ contains its supremum.

(b) T has elicitation complexity ∞ with respect to $(\mathcal{V}_k)_{k\in\mathbb{N}}$ unless T is constant.

Proof. Assume there is a $k \in \mathbb{N}$, a surjective functional $T' : \mathcal{F} \to \mathsf{A}'$ in $\mathcal{U}_k$ or $\mathcal{V}_k$ and a function $f : \mathsf{A}' \to \mathbb{R}$ such that $T = f \circ T'$. Without loss of generality, T' is surjective, hence its mixture-continuity together with the assumed convexity of $\mathcal{F}$ imply that A' is path-connected. Since it has non-empty interior, we can choose a hyperrectangle $Q := \prod_{i=1}^{k} [c_i, d_i] \subseteq \operatorname{int}(\mathsf{A}')$ and consider each component of T' isolated on Q. To do so, choose a component $j \in \{1, \dots, k\}$ and a $z_i \in [c_i, d_i]$ for all $i \in \{1, \dots, k\} \backslash \{j\}$. We can then obtain $F_{c_j,z}, F_{d_j,z} \in \mathcal{F}$ such that

$$\begin{aligned} T'(F_{c_j,z}) &= (z_1, \dots, z_{j-1}, c_j, z_{j+1}, \dots, z_k) \quad \text{and} \\ T'(F_{d_j,z}) &= (z_1, \dots, z_{j-1}, d_j, z_{j+1}, \dots, z_k). \end{aligned}$$

All components of T' are elicitable and thus have convex level sets by Theorem 2.3. Consequently, the i-th component, where $i \in \{1, \dots, k\} \backslash \{j\}$, equals z_i

for all convex combinations of $F_{c_j,z}$ and $F_{d_j,z}$. If we define

$$\mathsf{A}'_{j,z} := \{(z_1, \dots, z_{j-1}, x, z_{j+1}, \dots, z_k) \mid x \in (c_j, d_j)\} \subseteq Q,$$

the fact that the j-th component has convex level sets and is mixture-continuous implies that for all $a \in \mathsf{A}'_{j,z}$ there exists a $\lambda \in (0,1)$ with $T'(\lambda F_{c_j,z} + (1-\lambda)F_{d_j,z}) = a$. The connection $T = f \circ T'$ now gives

$$\begin{aligned} f((z_1, \dots, z_{j-1}, x, z_{j+1}, \dots, z_k)) &= f(T'(\lambda F_{c_j,z} + (1-\lambda)F_{d_j,z}) \\ &= T(\lambda F_{c_j,z} + (1-\lambda)F_{d_j,z}) \\ &= \max(T(F_{c_j,z}), T(F_{d_j,z})), \end{aligned}$$

for all $x \in (c_j, d_j)$, implying that f has to be constant on the set $\mathsf{A}'_{j,z}$. Repeating this argument for any choice of $j \in \{1, \dots, k\}$ and $z_i \in [c_i, d_i]$ with $i \in \{1, \dots, k\} \backslash \{j\}$ shows that there is a $C \in \mathbb{R}$ such that $f(q) = C$ for all $q \in \operatorname{int}(Q)$.

Now fix $x_0 \in \operatorname{int}(Q)$. For any $x_1 \in \mathsf{A}'$ we can choose distributions $F_0, F_1 \in \mathcal{F}$ with $T'(F_0) = x_0$ and $T'(F_1) = x_1$. Since $x_0 \in \operatorname{int}(Q)$ and T' is mixture-continuous, there is a small $\mu \in (0,1)$ such that $T'(\mu F_1 + (1-\mu)F_0) \in \operatorname{int}(Q)$ holds. We thus obtain

$$\begin{aligned} C = f(T'(\mu F_1 + (1-\mu)F_0)) &= T(\mu F_1 + (1-\mu)F_0) \\ &= \max(T(F_0), T(F_1)) \\ &= \max(f(x_0), f(x_1)) = \max(C, f(x_1)), \end{aligned}$$

implying $f(x_1) \leq C$. Since x_1 was arbitrary, we have $f(x) \leq C$ for all $x \in \mathsf{A}'$, showing $C = \sup T(\mathcal{F})$ and proving statement (a).

Assume now that A' is open. Then for every $x_1 \in \mathsf{A}'$ there is a hyperrectangle $Q_1 \subseteq \mathsf{A}'$ such that $x_1 \in \operatorname{int}(Q_1)$. Arguing as in the beginning of the proof gives $f(q) = f(x_1)$ for all $q \in \operatorname{int}(Q_1)$. So letting $T'(F_1) = x_1$ as above we obtain a $\nu \in (0,1)$ such that $T'(\nu F_1 + (1-\nu)F_0) \in \operatorname{int}(Q_1)$. This implies

$$\begin{aligned} C = f(T'(\mu F_1 + (1-\mu)F_0)) &= \max(T(F_0), T(F_1)) \\ &= f(T'(\nu F_1 + (1-\nu)F_0)) = f(x_1). \end{aligned}$$

Since x_1 was arbitrary, T must be constant, proving part (b). □

Theorem 3.7 implies infinite elicitation complexity of max-functionals in a wide range of natural settings. Ultimately, our main interest lies in understanding the elicitation complexity with respect to the more general family $\mathcal{U}_k$, which imposes only very weak assumptions on a potential intermediate functionals.

Corollary 3.8. *Let $T : \mathcal{F} \to \mathbb{R}$ be a max-functional and let one of the following conditions be satisfied.*

(i) T *is unbounded.*
(ii) T *is surjective onto an open interval* (a, b).
(iii) T *is surjective onto a half-open interval* $[a, b)$.

Then T *has elicitation complexity* ∞ *with respect to* $(\mathcal{U}_k)_{k\in\mathbb{N}}$.

Alternatively, considering elicitation complexity with respect to the family $(\mathcal{V}_k)_{k\in\mathbb{N}}$ amounts to requiring more regularity for a potential intermediate functional T' and, in this case, *all* non-constant max-functionals have infinite elicitation complexity. Lemma 2.7 further implies that the infinite elicitation complexity of max-functionals also extends to larger classes than the considered convex family of distribution functions $\mathcal{F}$ and is valid with respect to smaller families contained in $(\mathcal{U}_k)_{k\in\mathbb{N}}$ or $(\mathcal{V}_k)_{k\in\mathbb{N}}$.

Finally, by definition, any functional of finite elicitation complexity is conditionally elicitable, but it is unclear whether the reverse implication holds. We thus conclude with showing that max-functionals with infinite elicitation complexity can neither be conditionally elicitable nor jointly elicitable.

Theorem 3.9. *Let* $T : \mathcal{F} \to \mathbb{R}$ *be a max-functional such that* $\mathsf{elic}(T) = \infty$ *with respect to a family* $(\mathcal{C}_k)_{k\in\mathbb{N}}$. *Let* $T' : \mathcal{F} \to \mathsf{A}'$ *be a functional with* $T' \in \mathcal{C}_m$ *for some* $m \in \mathbb{N}$. *Then the following hold true.*

(a) T *is not conditionally elicitable given* T'.
(b) T *is not jointly elicitable with* T'.

Proof. For the first part assume conversely, that there is an $m \in \mathbb{N}$ and a functional $T' \in \mathcal{C}_m$ such that T is conditionally elicitable given T'. That is, T is elicitable on the subclass $\mathcal{F}_x = \{F \in \mathcal{F} \mid T'(F) = x\}$ for any $x \in \mathsf{A}'$. By assumption, there is no link function f such that $T = f \circ T'$ holds. Consequently, there is at least one $z \in \mathsf{A}' \subseteq \mathbb{R}^m$ such that T is not constant on $\mathcal{F}_z$. If z defines such a class, then it is convex due to the elicitability of T' and moreover we can find $F_0, F_1 \in \mathcal{F}_z$ such that $T(F_0) \neq T(F_1)$ holds. Theorem 3.3 now implies that the restriction of T to $\mathcal{F}_z$ cannot be elicitable, a contradiction to the conditional elicitability of T.

For the second part note that, as remarked in Section 2 and in the discussion of Fissler and Ziegel (2016), the joint elicitability of T with an elicitable functional T' implies that T is conditionally elicitable given T'. Consequently, the first part of the proof implies the result. □

We conclude this section with a technical remark. In the spirit of Frongillo and Kash (2018), our complexity result (Theorem 3.7) employs regularity assumptions on the possible intermediate functionals. The main assumption is that they possess elicitable components. Why this is essential is illustrated by the use of the hyperrectangle Q in the proof. Intuitively, this assumption can be relaxed at the cost of more technical arguments. The main challenge hereby is to control the values of T' in a small hyperrectangle (or ball) around some $x_0 \in \mathrm{int}(\mathsf{A}')$. However, we did not pursue this approach further, since we believe that our setting covers many functionals of practical interest and at the same

time illustrates the irregular behaviour that will be inherent to any link function for a max-functional.

4. Examples of max-functionals

Prominent examples of max-functionals, to which the results of Section 3 apply, are routinely considered in extreme value theory and are key characteristics for the purpose of inference on the tail of a distribution.

Upper endpoint For a real-valued random variable with distribution function F, its upper endpoint is the supremum of its support

$$x^F := \sup\{x \in \mathbb{R} \mid F(x) < 1\}.$$

By definition, the upper endpoint can be interpreted as a real-valued max-functional on the convex class $\{F \in \mathcal{F} \mid x^F < \infty\}$. Bellini and Bignozzi (2015, Example 3.9) discuss the upper endpoint under the name *worst-case risk measure* and show that it is not elicitable, once further regularity conditions on the admissible scoring functions are imposed. In light of Corollary 3.4 the non-elicitability of the upper endpoint follows without any further assumptions. In addition it has infinite elicitation complexity in the sense of Theorem 3.7 and Corollary 3.8.

Index of regular variation/Tail index When the upper endpoint is infinite, another key characteristic to describe the tail behaviour of heavy-tailed distributions is the index of regular variation. A strictly positive measurable function f satisfying

$$\lim_{x\to\infty} \frac{f(xt)}{f(x)} = t^\rho$$

for $t > 0$ is called *regularly varying (at infinity) with index* $\rho(f) \in \mathbb{R}$. For a distribution F its index of regular variation is the respective index for its survival function $\overline{F} := 1 - F$, that is, $T(F) := \rho(\overline{F})$. Its inverse $T(F)^{-1}$ is also called *tail index* in the risk management literature, cf. McNeil, Frey and Embrechts (2015, Section 5.1). If the tail $\overline{F}$ is regularly varying with (a negative) index ρ, this means that $\overline{F}$ decays essentially like a power function with decay rate $1/\rho$. Since $\rho(f + g) = \max(\rho(f), \rho(g))$ (cf. e.g. de Haan and Ferreira (2006, Proposition B.1.9)), the index of regular variation T is naturally a max-functional, while the tail index T^{-1} is a min-functional.

Tail-separating functionals More generally, we can deduce that the property of *'being a max-functional' (or min-functional)* is in fact inherent to all *'tail-ordering indices'*. To make this precise, let us consider the following natural order on distribution tails. For two distribution functions F and G with

upper endpoints $x^F, x^G \in \mathbb{R} \cup \{\infty\}$ we say that G *has heavier tail than* F and write $F <_t G$ if

$$\text{either} \quad x^F < x^G \qquad \text{or} \qquad x^F = x^G = x^* \text{ and } \lim_{x \to x^*} \frac{\overline{F}(x)}{\overline{G}(x)} = 0.$$

We say that F and G are *tail equivalent* and write $F \sim_t G$ if they share the same upper endpoint $x^F = x^G = x^* \in \mathbb{R} \cup \{\infty\}$ and

$$\lim_{x \to x^*} \frac{\overline{F}(x)}{\overline{G}(x)} \in (0, \infty).$$

Note that "$<_t$" defines a strict partial order on any set of distribution functions $\mathcal{F}$ and that for tail equivalent F and G neither $F <_t G$ nor $G <_t F$ can hold. The following proposition shows that a functional which respects the tail order "$<_t$" is a max-functional.

Proposition 4.1. *Let $T : \mathcal{F} \to \mathbb{R}$ be a functional that satisfies for all $F, G \in \mathcal{F}$*

$$T(F) - T(G) \begin{cases} \leq 0 & \text{if } F <_t G, \\ \geq 0 & \text{if } G <_t F, \\ = 0 & \text{else.} \end{cases}$$

Then T is a max-functional.

Proof. Let $F_0, F_1 \in \mathcal{F}$ and set $F_\lambda := \lambda F_1 + (1-\lambda) F_0$ for $\lambda \in (0,1)$. We distinguish three cases. If $F_0 <_t F_1$, we have $x^{F_\lambda} = x^{F_1} \geq x^{F_0}$ and the identity

$$\frac{\overline{F}_\lambda(x)}{\overline{F}_1(x)} = \lambda + (1-\lambda) \frac{\overline{F}_0(x)}{\overline{F}_1(x)}$$

for $x < x^{F_1}$ implies $F_\lambda \sim_t F_1$. Hence, neither $F_\lambda <_t F_1$ nor $F_1 <_t F_\lambda$ can be true. Together with $T(F_0) \leq T(F_1)$ we may conclude $T(F_\lambda) = T(F_1) = \max(T(F_0), T(F_1))$. By symmetry, the case $F_1 <_t F_0$ can be treated analogously. In the remaining case we have neither $F_0 <_t F_1$ nor $F_1 <_t F_0$, so $x^{F_1} = x^{F_0} = x^{F_\lambda} =: x^*$ must hold. Consequently,

$$\liminf_{x \to x^*} \frac{\overline{F}_\lambda(x)}{\overline{F}_1(x)} \geq \lambda > 0 \quad \text{and} \quad \limsup_{x \to x^*} \frac{\overline{F}_\lambda(x)}{\overline{F}_1(x)} < \infty,$$

where the latter follows as the tail of F_0 is not heavier than the tail of F_1. This implies that neither $F_1 <_t F_\lambda$ nor $F_\lambda <_t F_1$ can hold true, which gives $T(F_\lambda) = T(F_1) = \max(T(F_0), T(F_1))$ and concludes the proof. □

Another instance of a tail-ordering functional in the sense of Proposition 4.1 is the $\mathcal{M}$-*index* as introduced in Cadena and Kratz (2016). If it exists, it is the unique $\rho \in \mathbb{R}$ such that

$$\lim_{x \to \infty} \frac{\overline{F}(x)}{x^{\rho+\varepsilon}} = 0 \quad \text{and} \quad \lim_{x \to \infty} \frac{\overline{F}(x)}{x^{\rho-\varepsilon}} = \infty \quad \text{for all } \varepsilon > 0.$$

It is easily seen that the $\mathcal{M}$-index coincides with the index of regular variation for distribution functions F with regularly varying tail function $\overline{F}$. As it sorts survival functions according to their power law decay, Proposition 4.1 implies that the $\mathcal{M}$-index is a max-functional.

Extreme value index A central characteristics of extreme value theory is the extreme value index, which classifies the limiting behaviour of rescaled maxima of growing samples from a distribution. More precisely, if there exist suitable location-scale normings $a_n > 0$, $b_n \in \mathbb{R}$ such that the distribution functions $F_n(x) := F^n(a_n x + b_n)$ converge weakly to a non-degenerate distribution function G, the limiting distribution function G is necessarily a *Generalized Extreme Value Distribution (GEV)*. This means that up to a location-scale normalization we have

$$G(x) = G_\gamma(x) = \exp\{-(1+\gamma x)_+^{-1/\gamma}\}$$

for some $\gamma = \gamma(F) \in \mathbb{R}$, where $G_0(x) = \exp\{-e^{-x}\}$ for $\gamma = 0$. The distribution F is said to be in the *max-domain of attraction* of $G = G_\gamma$ and the shape parameter $\gamma(F)$ is the *extreme value index (EVI) of* F, cf. e.g. the monographs Resnick (1987) and de Haan and Ferreira (2006) for further background.

Let $\mathcal{F}$ be the class of distribution functions which are in a max-domain of attraction for some GEV and consider first the EVI on the subclass of heavy-tailed distributions $\mathcal{F}_+ = \{F \in \mathcal{F} \mid \gamma(F) > 0\}$. It is well-known that a distribution $F \in \mathcal{F}$ has EVI $\gamma > 0$ if and only if $\rho(\overline{F}) = -\gamma^{-1}$, where ρ is the index of regular variation (cf. e.g. Resnick (1987, Proposition 1.11)). Consequently, the EVI γ is also a max-functional on $\mathcal{F}_+$.

When considering the class of light-tailed distributions, i.e. the case $\gamma(F) < 0$, we need to specify an upper endpoint first in order to make 'being a max/min-functional' meaningful for the EVI γ. To this end, let $\mathcal{F}_{x^*} = \{F \in \mathcal{F} \mid \gamma(F) < 0,\ x^F = x^*\}$. Again the EVI behaviour is governed by regular variation, since $\gamma(F) = -\gamma(F_*)$ with $F_*(x) = F(x^* - x^{-1})$ (cf. e.g. Resnick (1987, Proposition 1.13)). This shows that the EVI γ is a min-functional on the class $\mathcal{F}_{x^*}$. Note that it is crucial to assume equal upper endpoints, because otherwise it is not the EVI that dominates the tail behaviour, but the upper endpoint itself.

So far, we have looked at statistical indices that classify *univariate* tail behaviour. However, similar issues arise when we want to quantify *joint* tail behaviour in higher dimensions. Exemplary, let us consider the coefficient of tail dependence.

Coefficient of tail dependence In order to quantify the tail behaviour of a bivariate distribution function Ledford and Tawn (1996, 1997) introduced the coefficient of tail dependence. For a bivariate distribution function F of a random vector (X_1, X_2) let us write $\overline{F}_i(x) := \mathbb{P}(X_i > x)$, $i = 1, 2$ and $\overline{F}(x) := \mathbb{P}(X_1 > x, X_2 > x)$ for the associated survival functions. Suppose there is an $\alpha > 0$ such that both $\overline{F}_1$ and $\overline{F}_2$ are regularly varying with index $-\alpha$. If in addition the joint survival function $\overline{F}$ is regularly varying with index $-\alpha/\eta$ for

some $\eta \in (0,1]$, the coefficient $\eta = \eta(F)$ is called *coefficient of tail dependence (CTD)* of the bivariate distribution F. Let us consider the CTD η on the class of bivariate distributions

$$\mathcal{F}_\alpha = \{F \mid \rho(\overline{F}_1) = \rho(\overline{F}_2) = -\alpha,\ \rho(\overline{F}) = -\alpha/\eta \text{ for some } \eta \in (0,1]\}.$$

Then it follows for $F, G \in \mathcal{F}_\alpha$ that $\rho(\lambda\overline{F} + (1-\lambda)\overline{G}) = -\alpha/\max(\eta(F), \eta(G))$ by the properties of the index of regular variation. Hence η is a max-functional on $\mathcal{F}_\alpha$.

5. Proper scoring rules and max-functionals

In probabilistic forecasting, the whole distribution function instead of a single value is reported to the decision maker. Analogously to a scoring function, a *scoring rule* then assigns a score based on the forecasted distribution and a realizing observation. The scoring rule is called *proper* if its expected score with respect to a distribution is minimized whenever the forecast coincides with this distribution, see e.g. Gneiting and Raftery (2007) or Dawid (2007) for recent reviews.

In light of the results of Section 3, the following approach may seem reasonable to someone seeking information about a max-functional: Instead of single values, distribution functions are reported and evaluated via proper scoring rules. Then the max-functionals are computed from the forecasted distributions.

If the max-functional of interest is a property of the tail, e.g. the extreme value index, one could expect this method to work well as long as the scoring rule shows a good performance in the tails. In order to emphasize specific regions of interests, in particular the tails, Gneiting and Ranjan (2011) and Diks, Panchenko and van Dijk (2011) combined scoring rules with weight functions. Drawbacks and benefits of these weighted proper scoring rules were further studied in Lerch et al. (2017) and Holzmann and Klar (2017), where the latter propose general construction principles. A theoretical problem is pointed out by Taillardat et al. (2019), who show that weighted versions of the continuously ranked probability score (CRPS) cannot detect that two distributions are not tail equivalent.

This section shows that the problems detected by Taillardat et al. (2019) occur also for max-functionals and do not depend on the specific choice of proper scoring rule. Simply put, the expected score difference of two distributions can be arbitrarily small while their values for a max-functional can be large. As previously, $\mathcal{F}$ is a convex set of distribution functions on $\mathsf{O} \subseteq \mathbb{R}^d$. In our notation we follow Gneiting and Raftery (2007) as well as Section 4.

Definition 5.1 (Scoring rule)**.** A real-valued function $S : \mathcal{F} \times \mathsf{O} \to \mathbb{R}$ is called *scoring rule* if for all $F \in \mathcal{F}$ the mapping $y \mapsto S(F, y)$ is $\mathcal{F}$-integrable. The scoring rule S is called *proper* if $\bar{S}(F, F) \leq \bar{S}(G, F)$ holds for all $F, G \in \mathcal{F}$. It is *strictly proper* if it is proper and for any $F, G \in \mathcal{F}$ the equality $\bar{S}(G, G) = \bar{S}(F, G)$ implies $G = F$.

For clarity of presentation we require all scoring rules to be $\mathcal{F}$-integrable, while Gneiting and Raftery (2007) only require *quasi-integrability*. The latter means that the expected score $\bar{S}(G, F)$ is well-defined (and not necessarily finite) for all $G, F \in \mathcal{F}$. Our assumption of $\mathcal{F}$-integrability is however only a minor restriction, which can be relaxed as discussed below.

A popular choice of scoring rule is the *(weighted) continuous ranked probability score*, abbreviated by *CRPS (wCRPS)*. For some weight function $w : \mathbb{R} \to [0, \infty)$ the wCRPS is defined via

$$\mathrm{wCRPS}(F, y) = \int_{-\infty}^{\infty} w(x)(F(x) - \mathbb{1}(y \leq x))^2 \, \mathrm{d}x$$

and the CRPS is obtained in the special case, where w is equal to one (Matheson and Winkler, 1976; Gneiting and Ranjan, 2011). In order to emphasize the right tail, the choice $w(x) = \mathbb{1}(q \leq x)$ for some threshold $q \in \mathbb{R}$ can be used. Both wCRPS and CRPS are proper scoring rules as long as $\mathcal{F}$ contains only distributions with finite first moments. In this case the CRPS is even strictly proper, while the wCRPS is only under additional assumptions, see Gneiting and Raftery (2007), Gneiting and Ranjan (2011) and Holzmann and Klar (2017).

As demonstrated by Taillardat et al. (2019, Section 2), the wCRPS is not able to clearly distinguish between different tail behavior. More precisely, given a distribution G and $\varepsilon > 0$, it is always possible to construct a distribution F that is not tail equivalent to G and such that

$$|\mathbb{E}\,\mathrm{wCRPS}(G, Y) - \mathbb{E}\,\mathrm{wCRPS}(F, Y)| \leq \varepsilon,$$

where Y has distribution G. This results shows that for any distribution G the tail can be modified while keeping the expected wCRPS ε-close to its minimum. As put by Taillardat et al. (2019) this means that the wCRPS is not a *tail equivalent score*.

In the following we show that *all* proper scoring rules fail to be tail equivalent in this sense. Moreover, we extend these findings to max-functionals, i.e. we show that no proper scoring rule is *max-functional equivalent*. Both findings are immediate consequences of the subsequent continuity considerations for scoring rules.

Definition 5.2. A scoring rule $S : \mathcal{F} \times \mathsf{O} \to \mathbb{R}$ is called *diagonal-continuous at* G if for all $F \in \mathcal{F}$

$$\bar{S}(\lambda F + (1 - \lambda)G, G) \to \bar{S}(G, G) \qquad \text{for } \lambda \downarrow 0.$$

Lemma 5.3. *If $S : \mathcal{F} \times \mathsf{O} \to \mathbb{R}$ is a proper scoring rule, it is diagonal-continuous at each $G \in \mathcal{F}$.*

Proof. We proceed similar to the proof of Nau (1985, Proposition 3). Let $F, G \in \mathcal{F}$ and denote $F_\lambda := \lambda F + (1 - \lambda)G$ for $\lambda \in [0, 1)$. We obtain the inequality

$$(1 - \lambda)\bar{S}(F_\lambda, G) = \bar{S}(F_\lambda, F_\lambda) - \lambda \bar{S}(F_\lambda, F)$$

$$\leq \bar{S}(G, F_\lambda) - \lambda \bar{S}(F, F)$$
$$= (1-\lambda)\bar{S}(G, G) + \lambda\big(\bar{S}(G, F) - \bar{S}(F, F)\big),$$

since S is a proper scoring rule. Rearranging leads to

$$|\bar{S}(\lambda F + (1-\lambda)G, G) - \bar{S}(G, G)| \leq \frac{\lambda}{1-\lambda}\big(\bar{S}(G, F) - \bar{S}(F, F)\big),$$

for $\lambda \in [0, 1)$ and the right hand side of this equation vanishes as $\lambda \downarrow 0$. □

The argument of the proof of Lemma 5.3 can be extended to quasi-integrable scoring rules as considered in Gneiting and Raftery (2007). The additional requirement is that the expected score $\bar{S}(G, F)$ is finite and that S is *regular*, i.e. $\bar{S}(F, F) \in \mathbb{R}$ for all $F \in \mathcal{F}$.

We can now turn our attention to the main result of this section. It is motivated by the observation that tail equivalence and max-functionals lead to a similar kind of discontinuity on the convex combinations $\lambda F + (1-\lambda)G$, which intuitively conflicts with the diagonal-continuity of proper scoring rules. This allows for an extension of the results of Taillardat et al. (2019). Recall the tail-ordering from Section 4 and that we assume $\mathcal{F}$ to be convex.

Theorem 5.4. *Let $S : \mathcal{F} \times \mathbb{R} \to \mathbb{R}$ be a proper scoring rule and $G \in \mathcal{F}$. Then the following are true.*

(a) If there is an $F \in \mathcal{F}$ with heavier tail than G, then for all $\varepsilon > 0$ there is an $F_\varepsilon \in \mathcal{F}$ that is not tail equivalent to G and such that

$$|\bar{S}(F_\varepsilon, G) - \bar{S}(G, G)| \leq \varepsilon.$$

(b) Let $T : \mathcal{F} \to \mathbb{R}$ be a max-functional. If there is an $F \in \mathcal{F}$ with $T(F) > T(G)$, then for all $\varepsilon > 0$ there is an $F_\varepsilon \in \mathcal{F}$ such that $T(F_\varepsilon) = T(F) > T(G)$, while

$$|\bar{S}(F_\varepsilon, G) - \bar{S}(G, G)| \leq \varepsilon.$$

Proof. Fix $G \in \mathcal{F}$ and let S be a proper scoring rule. For $F \in \mathcal{F}$ set $F_\lambda := \lambda F + (1-\lambda)G$. Since $\mathcal{F}$ is convex, we have $F_\lambda \in \mathcal{F}$ for all $\lambda \in [0, 1]$. Moreover, S is diagonal-continuous at G by Lemma 5.3, implying that for all $\varepsilon > 0$ and $F \in \mathcal{F}$ we can find a $\delta \in (0, 1]$ such that $|\bar{S}(F_\lambda, G) - \bar{S}(G, G)| \leq \varepsilon$ holds for all $\lambda \in [0, \delta]$. Now assume there is an $F \in \mathcal{F}$ with heavier tail than G. If $x^F > x^G$, we have $x^{F_\lambda} > x^G$ for all $\lambda \in (0, 1]$. If on the other hand $x^F = x^G = x^*$ we have

$$\frac{\overline{F}_\lambda(x)}{\overline{G}(x)} = (1-\lambda) + \lambda\frac{\overline{F}(x)}{\overline{G}(x)},$$

for $x < x^*$ and the right-hand side goes to infinity as $x \to x^*$. Hence, in both cases the distributions F_λ cannot be tail equivalent to G for $\lambda \in (0, 1]$, showing part (a). For the second part, let $F \in \mathcal{F}$ satisfy $T(F) > T(G)$. Since T is a max-functional, $T(F_\lambda) = T(F) > T(G)$ holds for $\lambda \in (0, 1]$, proving part (b). □

The first part of Theorem 5.4 shows that the lack of tail equivalence is not a flaw of the wCRPS, but inherent to *all* proper scoring rules (up to integrability assumptions). The second part extends this non-equivalence of proper scoring rules to max-functionals. Loosely speaking, this means that there can not only be pairs of not tail equivalent distributions, but also pairs of distributions with arbitrarily different max-functional values, and both having almost identical expected scores.

6. Discussion

Recent research investigates the elicitation properties of widely used statistical functionals. When the emphasis lies on an understanding of tail properties, typical functionals to characterize this behaviour fall into the class of max-functionals. In particular, all functionals that order distribution tails belong to this class (cf. Proposition 4.1). We show here that max-functionals do not only fail to be elicitable (Theorem 3.3), but have in fact infinite elicitation complexity in a wide range of settings (Theorem 3.7). This contrasts situations in which the non-elicitability can be alleviated by a finite elicitation complexity as, for instance, is the case for the variance or the Expected Shortfall (Frongillo and Kash, 2018; Fissler and Ziegel, 2016). Rather it bears resemblance to the mode, which is non-elicitable and has infinite elicitation complexity as well, see Heinrich (2014) and Dearborn and Frongillo (2019). As an alternative to point forecasts, we may allow the max-functional to be reported via the entire distribution function. In principle such probabilistic forecasts can be compared using proper scoring rules. However, Theorem 5.4 demonstrates that the difference of expected scores can be arbitrarily small, although the difference of max-functional values may be large. The latter complements recent findings of Taillardat et al. (2019) and extends them from the wCRPS to all integrable proper scoring rules.

Collectively, our results cast doubt on the ability of expected scores to distinguish different tail regimes in the sense of max-functional values as they are routinely considered in extreme value theory. From an applied viewpoint this means that expected scores are not suitable to access such tail information for regression, M-estimation or comparative forecast evaluation. Thereby, our results provide a new perspective on the limitations of weighted scoring rules, adding to practical intricacies described in Lerch et al. (2017), Holzmann and Klar (2017) and Friederichs and Thorarinsdottir (2012). What might come to rescue though, is that the max-functionals themselves are often not the main concern in applications, but rather a tool to guide the extrapolation from intermediate order statistics to the functionals of interest. In practice, these functionals may include a high quantile or a tail expectation such as Expected Shortfall, which can be interpreted as tail properties 'less extreme' than max-functionals and with better elicitablity properties.

Lastly, however, we would like to point out that non-elicitability is not the only problem in sound forecast evaluation and many open questions remain.

Even when elicitability is granted, i.e. the considered statistical functional is the unique minimizer of an expected score, there is no guarantee that the corresponding minimization problem will be well-posed. For instance, poorly behaved scoring functions may give rise to high variances of realized average scores, in which case practical sample sizes may be per se too low for an adequate assessment of competing forecasts. Due to the many challenges in forecast evaluation with an emphasis on distribution tails, we anticipate that it will remain an active area of research.

Acknowledgments

Jonas Brehmer gratefully acknowledges support by the German Research Foundation (DFG) through the Research Training Group RTG 1953. The authors would also like to thank Tilmann Gneiting, Fabian Krüger and Martin Schlather for their valuable comments.

References

Bellini, F. and Bignozzi, V. (2015). On elicitable risk measures. *Quantitative Finance* **15** 725–733. MR3334566

Cadena, M. and Kratz, M. (2016). New results for tails of probability distributions according to their asymptotic decay. *Statistics & Probability Letters* **109** 178–183. MR3434975

Dawid, A. P. (2007). The geometry of proper scoring rules. *Annals of the Institute of Statistical Mathematics* **59** 77–93. MR2396033

de Haan, L. and Ferreira, A. (2006). *Extreme value theory. Springer Series in Operations Research and Financial Engineering*. Springer, New York. MR2234156

Dearborn, K. and Frongillo, R. (2019). On the indirect elicitability of the mode and modal interval. *Annals of the Institute of Statistical Mathematics*. To appear. Available at https://doi.org/10.1007/s10463-019-00719-1.

Diks, C., Panchenko, V. and van Dijk, D. (2011). Likelihood-based scoring rules for comparing density forecasts in tails. *Journal of Econometrics* **163** 215–230. MR2812867

Emmer, S., Kratz, M. and Tasche, D. (2015). What is the best risk measure in practice? A comparison of standard measures. *Journal of Risk* **18** 31–60.

Ferro, C. A. T. and Stephenson, D. B. (2011). Extremal dependence indices: Improved verification measures for deterministic forecasts of rare binary events. *Weather and Forecasting* **26** 699–713.

Fissler, T., Hlavinová, J. and Rudloff, B. (2019). Elicitability and identifiability of systemic risk measures and other set-valued functionals. Available at https://arxiv.org/pdf/1907.01306.pdf.

Fissler, T. and Ziegel, J. F. (2015). Higher order elicitability and Osband's principle. Available at https://arxiv.org/pdf/1503.08123v3.pdf.

Fissler, T. and Ziegel, J. F. (2016). Higher order elicitability and Osband's principle. *The Annals of Statistics* **44** 1680–1707. MR3519937

Fissler, T. and Ziegel, J. F. (2019). Order-sensitivity and equivariance of scoring functions. *Electronic Journal of Statistics* **13** 1166–1211. MR3935847

Friederichs, P. and Thorarinsdottir, T. L. (2012). Forecast verification for extreme value distributions with an application to probabilistic peak wind prediction. *Environmetrics* **23** 579–594. MR3020076

Frongillo, R. and Kash, I. A. (2015). Vector-valued property elicitation. *Journal of Machine Learning Research: Workshop and Conference Proceedings* **40** 1–18.

Frongillo, R. and Kash, I. A. (2018). Elicitation complexity of statistical properties. Available at https://arxiv.org/pdf/1506.07212.pdf.

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association* **106** 746–762. MR2847988

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* **102** 359–378. MR2345548

Gneiting, T. and Ranjan, R. (2011). Comparing density forecasts using threshold- and quantile-weighted scoring rules. *Journal of Business & Economic Statistics* **29** 411–422. MR2848512

Heinrich, C. (2014). The mode functional is not elicitable. *Biometrika* **101** 245–251. MR3180670

Holzmann, H. and Klar, B. (2017). Focusing on regions of interest in forecast evaluation. *The Annals of Applied Statistics* **11** 2404–2431. MR3743302

Lambert, N. S., Pennock, D. M. and Shoham, Y. (2008). Eliciting properties of probability distributions. In *Proceedings of the 9th ACM Conference on Electronic Commerce. EC '08* 129–138. ACM.

Ledford, A. W. and Tawn, J. A. (1996). Statistics for near independence in multivariate extreme values. *Biometrika* **83** 169–187. MR1399163

Ledford, A. W. and Tawn, J. A. (1997). Modelling dependence within joint tail regions. *Journal of the Royal Statistical Society. Series B. Methodological* **59** 475–499. MR1440592

Lerch, S., Thorarinsdottir, T. L., Ravazzolo, F. and Gneiting, T. (2017). Forecaster's dilemma: extreme events and forecast evaluation. *Statistical Science. A Review Journal of the Institute of Mathematical Statistics* **32** 106–127. MR3634309

Matheson, J. E. and Winkler, R. L. (1976). Scoring rules for continuous probability distributions. *Management Science* **22** 1087–1096.

McNeil, A. J., Frey, R. and Embrechts, P. (2015). *Quantitative risk management*, revised ed. *Princeton Series in Finance*. Princeton University Press, Princeton, NJ. MR3445371

Nau, R. F. (1985). Should scoring rules be 'effective'? *Management Science* **31** 527–535.

Osband, K. (1985). Providing Incentives for Better Cost Forecasting, PhD thesis, University of California, Berkely.

Resnick, S. I. (1987). *Extreme values, regular variation, and point processes.*

Applied Probability. A Series of the Applied Probability Trust **4**. Springer-Verlag, New York. MR900810

Steinwart, I., Pasin, C., Williamson, R. and Zhang, S. (2014). Elicitation and identification of properties. *Journal of Machine Learning Research: Workshop and Conference Proceedings* **35** 1–45.

Stephenson, D. B., Casati, B., Ferro, C. A. T. and Wilson, C. A. (2008). The extreme dependency score: A non-vanishing measure for forecasts of rare events. *Meteorological Applications* **15** 41–50.

Taillardat, M., Fougères, A.-L., Naveau, P. and De Fondeville, R. (2019). Extreme events evaluation using CRPS distributions. Available at https://hal.archives-ouvertes.fr/hal-02121796/file/CRPS-190429.pdf.

Weber, S. (2006). Distribution-invariant risk measures, information, and dynamic consistency. *Mathematical Finance* **16** 419–441. MR2212272